

DOCUMENT RESUME

ED 120 238

95

TM 005 197

AUTHOR Anderson, Richard B.
TITLE Follow Through: Testing One Model of Evaluation and Several Models of Compensation.
INSTITUTION Abt Associates, Inc. Cambridge, Mass.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Apr 76
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976); Charts may be of marginal legibility due to poor reproducibility of original

EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage
DESCRIPTORS Achievement Gains; Achievement Tests; Comparative Analysis; *Compensatory Education Programs; Data Analysis; *Disadvantaged Youth; *Early Childhood Education; Intervention; Mathematics; National Norms; Program Effectiveness; *Program Evaluation; Reading; *Research Methodology; Spelling; Test Results; Validity

IDENTIFIERS Metropolitan Achievement Tests; *Project Follow Through

ABSTRACT

The Follow Through experience is beginning to provide compelling evidence bearing upon important policy issues in compensatory education. This evidence, coming as it does from a naturalistic quasi-experiment, cannot draw its plausibility from the conventional statistical inference appropriate to probabilistically-constructed true experiments. The latest Follow Through results illustrate how appropriate cross-validity analytic techniques can yield convincing patterns of accomplishment by diverse "sponsors" who have implemented their programs in various settings. (Author)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED120238

FOLLOW THROUGH:
TESTING ONE MODEL OF EVALUATION
AND SEVERAL MODELS OF COMPENSATION

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

by

Richard B. Anderson
Abt Associates Inc.

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Presented at the Annual Meeting of the American Educational
Research Association at San Francisco, 20 April 1976

The Follow Through experience is beginning to provide compelling evidence bearing upon important policy issues in compensatory education. This evidence, coming as it does from a naturalistic quasi-experiment, cannot draw its plausibility from the conventional statistical inference appropriate to probabilistically-constructed true experiments. The latest Follow Through results illustrate how appropriate cross-validating analytic techniques can yield convincing patterns of accomplishment by diverse "sponsors" who have implemented their programs in various settings.

This paper is based on research performed by Abt Associates Inc. under Contract No. 300-75-0134 to the United States Office of Education. This research is the product of many hands; neither Abt Associates staff nor the Office of Education is responsible for any errors or omissions: the author accepts full responsibility for those. The statements in this paper do not necessarily represent any official position of the Office of Education.

TM005 197

How can we get trustworthy answers to evaluative questions from the results of a naturalistic quasi-experiment? If we insist on unbiased estimates with known probabilities of being wrong to any specified degree, the answer has to be that we can't. Quasi-experiments, and any statistical computations we may make with their results, have no rigorous probabilistic basis. As estimates of the characteristics of other groups than those studied directly, the naturalistic sample's characteristics are generally biased to an unknown extent and often in an unknown direction.

On the other hand, the only evidence that we shall ever have bearing upon a number of important questions will have to come from naturalistic observations. Other questions that might ultimately be addressed by experiments have not been, thus far, and so we must either make use of the information we have or act in ignorance, as if we had none at all. Where, as in Follow Through, extensive evaluative data exist, the decision maker will clearly be wise to learn from those data, even if the lessons they teach are not "strictly scientific." Some observations are bound to emerge with practical certainty from any extended set of evaluative data; with luck, these may be among the most important ones for policy purposes. Depending on his level of tolerance for uncertainty, the policymaker can gain additional insights at the heuristic level. Still others, more fine-grained, will remain beyond his reach.

The analyst's task in the process of evaluation is to get from the data the trustworthy conclusions that are actually there without being deceived by the noise and biases that are also inevitably there. The more naturalistic a quasi-experiment is, the less does its design dictate a single "appropriate" analysis: several diverse but parallel analytic procedures generally reflect, each imperfectly, the logical content of the evaluative questions. The primary analyst therefore usually follows one approach, to which each reader then reacts: "but what if he had analyzed the data this way?" A series of secondary analyses then typically ensues, some supporting the results of the primary analysis, and others calling them into question.

Campbell and others (1974) and Cook (1974) have recommended strategies akin to analytic cross-validation (ACV) for quasi-experimental studies: under this strategy, when faced with a number of imperfect analytic alternatives, one chooses all of them and compares their results. In effect, the program of primary analysis followed by a series of secondary analyses is a stretched-out version of the analytic cross-validation procedure. Doing the primary and secondary analyses together probably maximizes their comparability at some cost in the benefits of hindsight and independence. Smith (1973) and Weisberg (1974) have applied a simultaneous cross-validation strategy in their evaluation of Follow Through's cousin, Head Start Planned Variation. The present application in the Follow Through national evaluation is the largest and most systematic that we know of to date.

Follow Through was designed in the middle 1960's as a sequel to Head Start, providing compensatory services to disadvantaged children from kindergarten through third grade. As an "experiment," Follow Through has sought to find out:

- whether any of a number of kinds of compensatory intervention has the power to eliminate educational disadvantage in young children;
- which interventions work better than others; and
- what determines the success or failure of an intervention.

The design called for a number of educational specialists with varying philosophies to serve as "sponsors": to develop educational approaches and to apply their diverse theoretical models in eligible school districts throughout the nation. The data presented here reflect selected aspects of the performance of thirteen of these sponsors; the analytic sample comprises 3,367 children in 52 sites across the nation. These sites were selected on the basis of eligibility and willingness to cooperate with sponsors in implementing compensatory programs. The pairing of sites and sponsors was thoroughly judgmental. Within each site, eligible schools were identified, and then non-Follow Through (NFT) comparison schools were selected judgmentally. Intact Follow Through (FT) classroom groups received the programs, and Follow Through and non-Follow Through groups were tested at entry to kindergarten and at the end of third grade. Some intermediate testing was conducted in the spring of each intervening year, but the test batteries

were restricted, and many sites were not tested. Parent interviews and teacher questionnaires were completed both at entry and at exit.

For the purposes of this paper, we shall focus on the reading, mathematics, and spelling subtest results of the Metropolitan Achievement Test administered at the end of third grade. The outcome test battery included a number of other academic and non-academic measures, and a comprehensive evaluation of this multifaceted program would require data on many other aspects of compensatory accomplishment, but these will suffice to illustrate our primarily methodological point. The interested reader is referred to Abt Associates' report entitled "Education as Experimentation: A Planned Variation Model, Volume 3" which reports the complete evaluation based on these data.

The battery of analyses in our analytic cross-validation scheme includes the following:

- To address the question, "Has Follow Through eliminated the disadvantage with which its subjects entered school?" we have compared the Follow Through children's scores with national norms, in two ways:
 - We present the national grade-equivalent deviation of each site-group's median posttest scores from "grade-level"; and
 - We present for each Follow Through site-group the percentage of children who complete their Follow Through experience one year or more behind the national grade-level norm.

We recognize, be it hastily interjected here, the many disadvantages of norm score comparisons for evaluative purposes. It is only in the context of analytic cross-validation that we would feel that we could conscientiously present such comparisons. In this context, they are particularly illuminating, as will appear later.

- To address the question, "Have Follow Through children done better than they would have done without Follow Through?" we have conducted three parallel analyses of adjusted outcomes, variations on the general linear model:
 - An analysis of covariance which compares Follow Through posttest scores with non-Follow Through scores in the same site, adjusting for pretest scores and various covariate indicators of initial disadvantage, such as income, ethnicity, and levels of parental education and occupation.

- An analysis of covariance which compares each Follow Through group with a non-Follow Through group matched to it in factors of disadvantage, but not located in the same site;
- A residuals analysis which predicts each child's posttest score based on his characteristics of disadvantage, his pretest, and his site membership, in accordance with a regression model developed on non-Follow Through children; children's predicted scores are compared with their actual scores, and the aggregated differences are interpreted as estimates of program effect.

Once again, group matching is a perilous procedure which could not defensibly stand alone as the basis for evaluative analysis. The local comparison groups, being judgmentally selected, are nearly as unreliable, but they at least provide a crude control for salient regional and community environment differences. The residuals analysis, being essentially a comparison of each Follow Through group with an overall pooled non-Follow Through, begins to shade conceptually toward the national norms analysis.

- Finally, we have performed a simple ranking analysis of the 52 Follow Through site-group scores, together with the 36 non-Follow Through groups associated with them. This analysis compares the rank of each site group's posttest score with that of its pretest score, among the 88 groups. A site which might rank 20th at pretest and 3rd at posttest would look good on this analysis.

We have presented the results of these analyses in two redundant but complementary forms. Figures 1a-1c illustrate the format of parallel analyses within sites within outcomes: they address the question, "Is there an effect here?" For each of the five sites in Sponsor E*, for example, Figure 1a displays the mean MAT Total Reading score for the Follow Through children in comparison to the national norm mean and to the overall mean of all non-Follow Through comparison children. It also shows, side by side, arrows depicting the results of the three adjusted gains analyses described above: an arrow pointing upward indicates a higher adjusted score for the Follow Through group than for its comparison (a "positive effect"); an arrow pointing downward indicates the opposite. Arrows with large heads indicate "effect" estimates larger than one-quarter of a standard deviation

* We suppress the identities of sponsors and sites in this primarily methodological paper. Readers interested in the evaluation of Follow Through will find full details in Volume 3 of the Abt Associates report, already cited.

Figure 1a

DIFFERENCES BETWEEN FT AND NFT ADJUSTED THIRD GRADE OUTCOME
ON MAT TOTAL READING
FOR COHORT II SITES
IN SPONSOR E

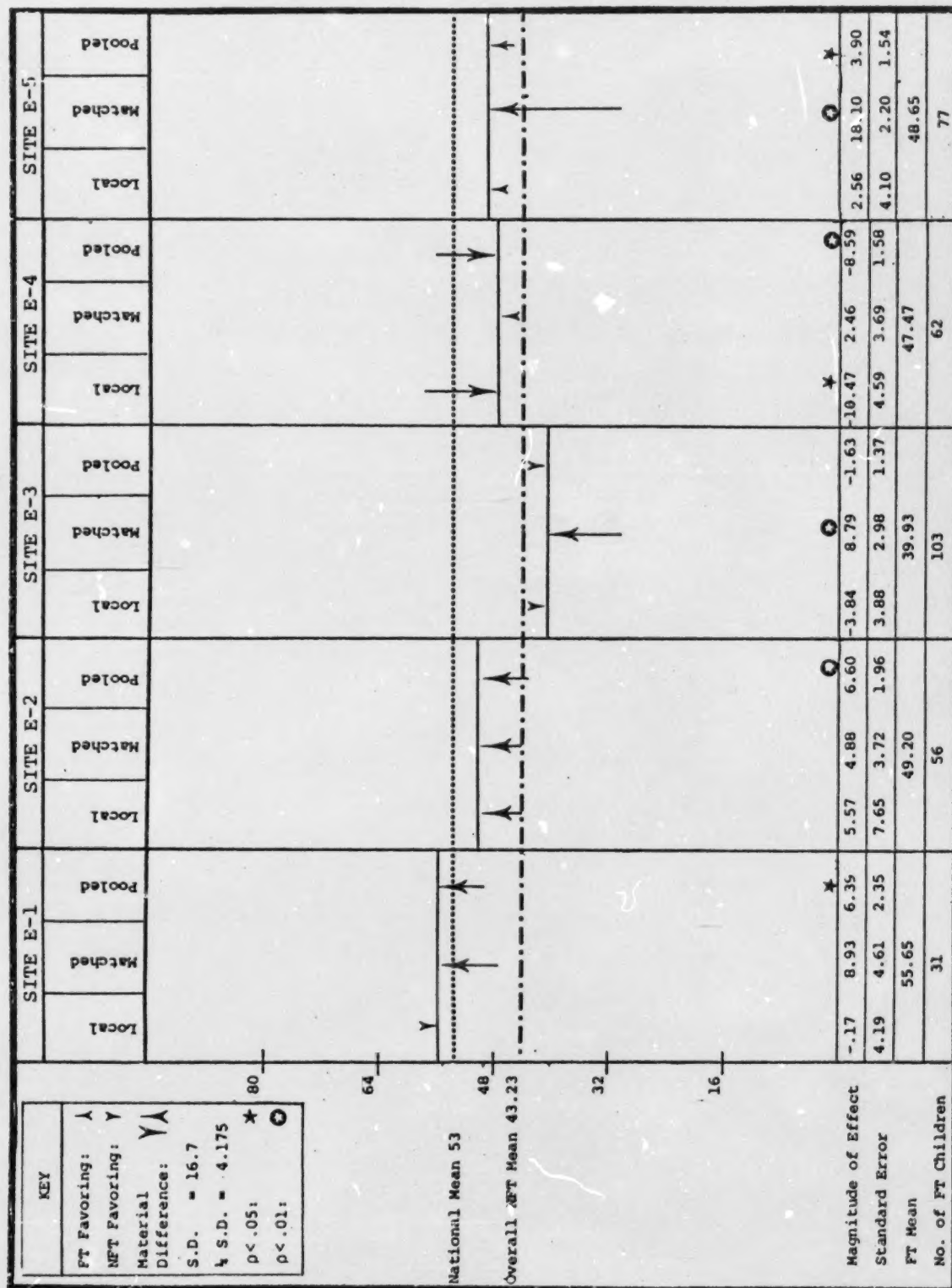


Figure 1b

DIFFERENCES BETWEEN FT AND NFT ADJUSTED THIRD GRADE OUTCOME
ON MAT TOTAL MATHEMATICS
FOR COHORT II SITES
IN SPONSOR E

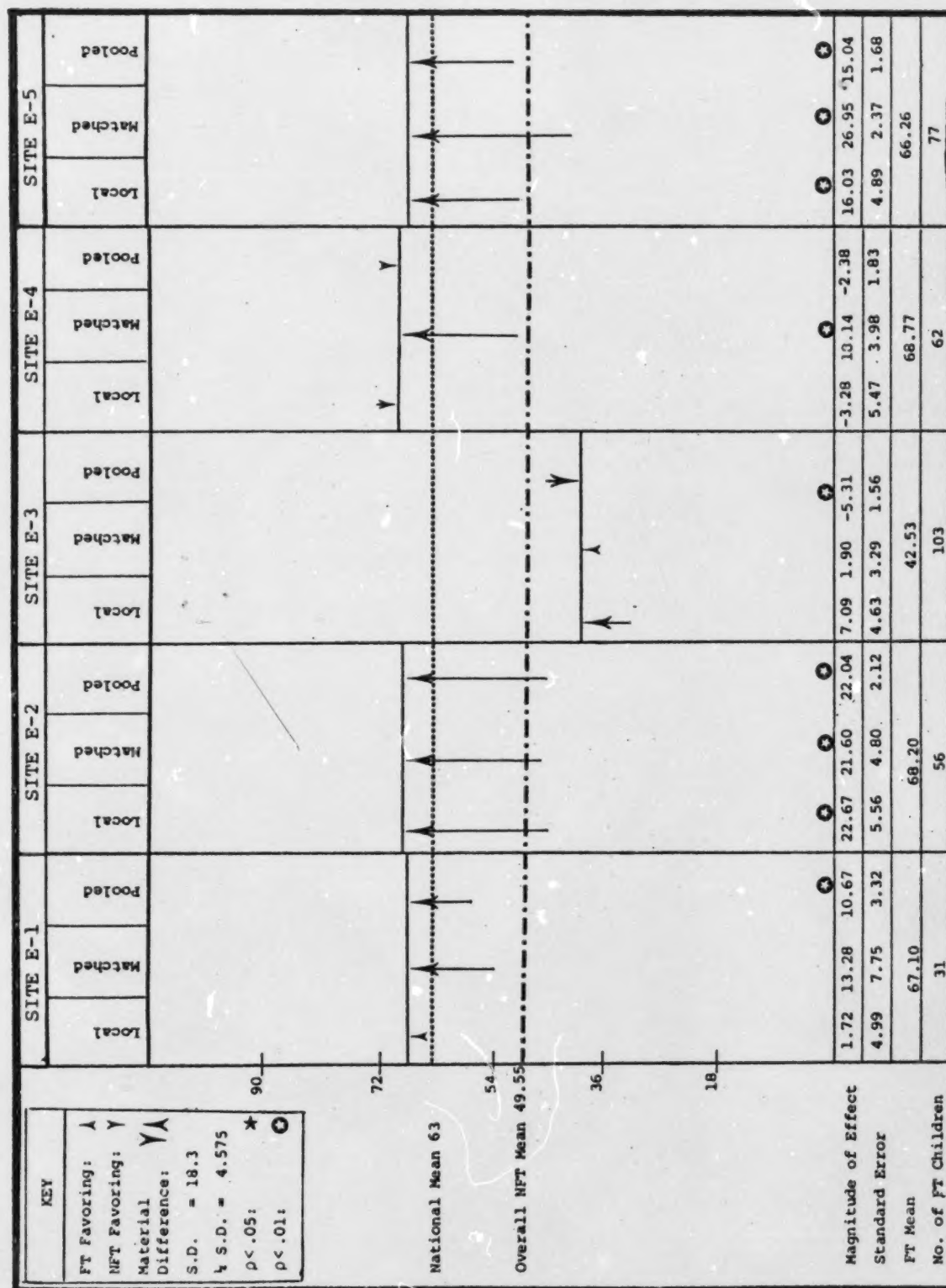
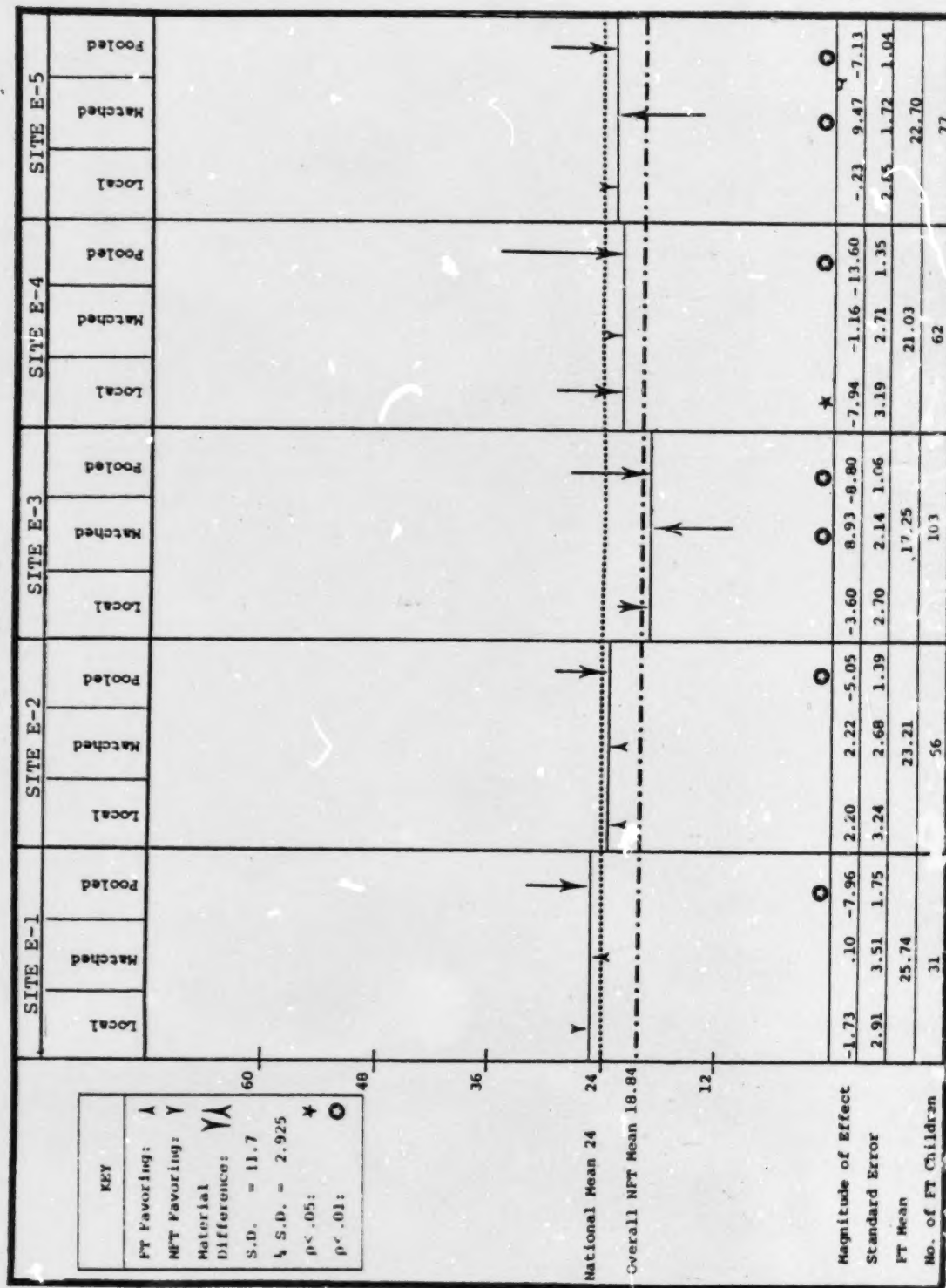


Figure 1c

DIFFERENCES BETWEEN FT AND NFT ADJUSTED THIRD GRADE OUTCOME
ON MAT SPELLING
FOR COHORT II SITES
IN SPONSOR E



of the outcome measure; we have referred to such an effect heuristically as a "material effect." Stars denote "significant" effects at two pseudo-probabilistic levels.

In Figure 1-b, Site E-2 illustrates both the sponsor's dream and that of the evaluator: all three analyses agree that the site's Follow Through children score some 22 points higher than those in the three comparison groups, once account has been taken of differential initial conditions. This difference, both "significant" and "material," places the children above both the NFT and the national means. If all the site results were like these, we would have both impressive and consistent effects estimates, with no inter-analytic ambiguities to explain. We would also quickly lose interest in analytic cross-validation, since any one analysis would tell us as much as all of them together.

In Figure 1-c, Site E-5 illustrates a more problematic situation: the FT group's adjusted performance is indistinguishable from that of its local comparison, better than that of its matched comparison, and worse than that predicted by the regression model based on all NFT children. The three analyses agree closely, however, in their estimates of the sponsor's effect on mathematics achievement in this same site (Figure 1-b). Lacking evidence on which to base a preference of one comparison over another, we judge this site's spelling effects ambiguous and so report them. If severe ambiguity in the interpretation of these parallel analyses were the rule rather than the exception, we should have reason to doubt the usefulness of the entire process--but not that of the analytic cross-validation strategy which would have revealed the ambiguity to us. Had we done only one analysis in such a situation, we would have the worst of all possible evaluative worlds: utterly inconclusive results together with ignorance of their inconclusiveness.

For the Follow Through data, effect estimates of interesting magnitude are, by and large, acceptably consistent from one analysis to another. Seldom, indeed, do the differences among them alter the gross picture on which we need to base our principal conclusions. Where discrepancies in more than detail do appear, a little digging into the detailed data usually suffices to learn why they arise and to illuminate their evaluative meaning. The

very presence of such discrepancies serves effectively to call the evaluator's attention to potentially important "wrinkles" in the data.

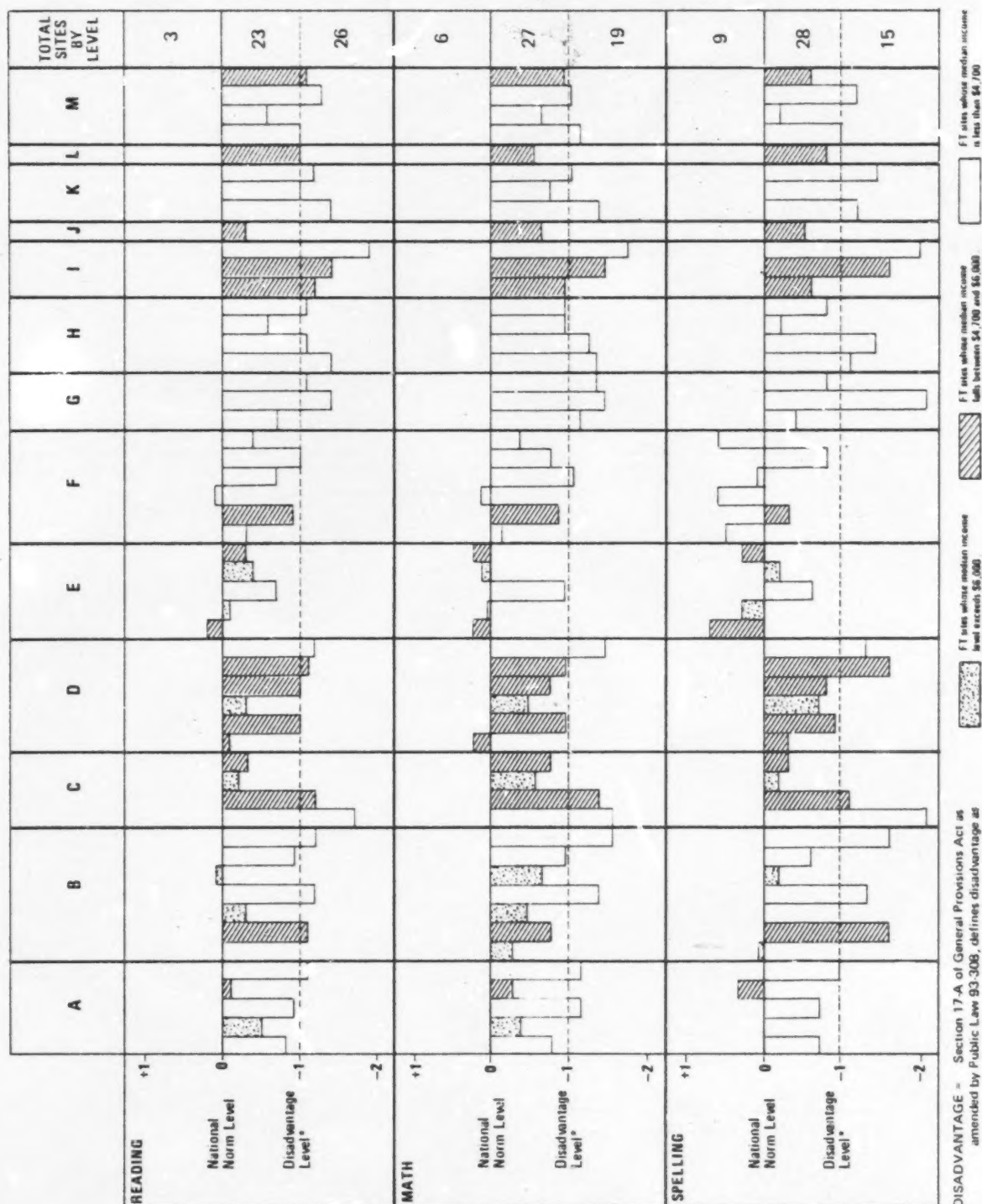
To illustrate this point, Figures 2a through 2f exemplify the format of parallel sites and outcomes within analyses. This format asks the question, "Are these analyses telling the same story?" Where they are, we conclude that we have squeezed from these data about as much as they contain. Where they are not, we look more deeply. Take, for example, the spelling scores for Sponsor K. Figure 2a tells us that the median child in each of this sponsor's three sites is around a grade behind the national norm in spelling skills. Figure 2b says that about half the children in each of these three sites (not surprisingly) fall below this legal definition of "educational disadvantage." The local comparisons for Sponsor K in Figure 2c are all small and negative; the best match comparison (Figure 2d) reverses the direction of one of the three. But on the pooled comparison (Figure 2e), Sponsor K nearly monopolizes the positive effects.

How can we interpret this contradiction? Must we merely say that we don't know what happened, or can other facts to be found in the data enable us to make evaluative sense of the situation? First, I note that this sponsor emphasizes the teaching of the conventions of the Spanish language to children whose native language is not standard English. Since the other sponsors generally consider language conventions to be a lower-priority objective, it would not be surprising to find Sponsor K outperforming them with respect to spelling. Secondly, we note that the local FT and NFT groups in the three sites are rather badly mismatched: the FT children are generally more disadvantaged and include a higher proportion of minorities -- especially linguistic minorities -- than the NFT children; the local ANCOVA is therefore immediately suspect. The best-matched sites, as it turns out, are hardly better matched in these substantively crucial respects; one of them, indeed, is the same as the local comparison, since the FT group was so distinctive that no better comparison group could be identified.

We therefore suspect that both the local and matched comparisons are telling less than the whole story. With treatment even partially confounded with an important covariable, we doubt whether that covariable can accomplish

Figure 2a

GRADE EQUIVALENCE DEVIATIONS FROM NATIONAL NORMS:
MEDIAN CHILD SCORES ON MAT READING, MATHEMATICS AND SPELLING
FOR COHORT II SITES



DISADVANTAGE = Section 17-A of General Provisions Act as amended by Public Law 93-308, defines disadvantage as achievement which is one or more years behind that expected at the appropriate grade level.

Figure 2b
PROPORTION OF CHILD "DISADVANTAGE" WITH RESPECT TO
MAT READING, MATHEMATICS AND SPELLING
FOR COHORT II SITES

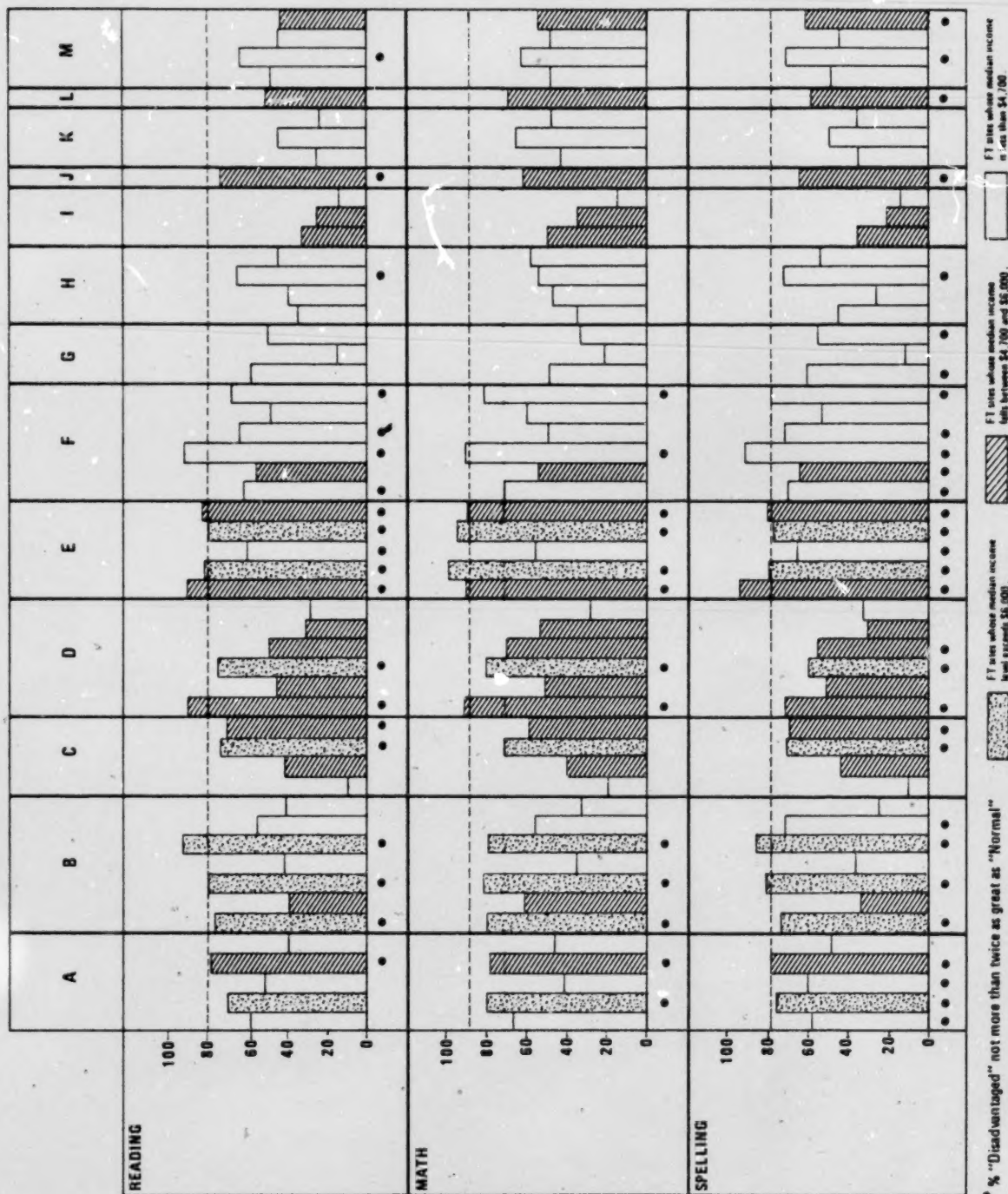


Figure 2c

DIFFERENCES BETWEEN FT AND LOCAL NFT ADJUSTED THIRD GRADE OUTCOMES
ON MAT READING, MATHEMATICS AND SPELLING IN STANDARD DEVIATION UNITS
FOR COHORT II SITES

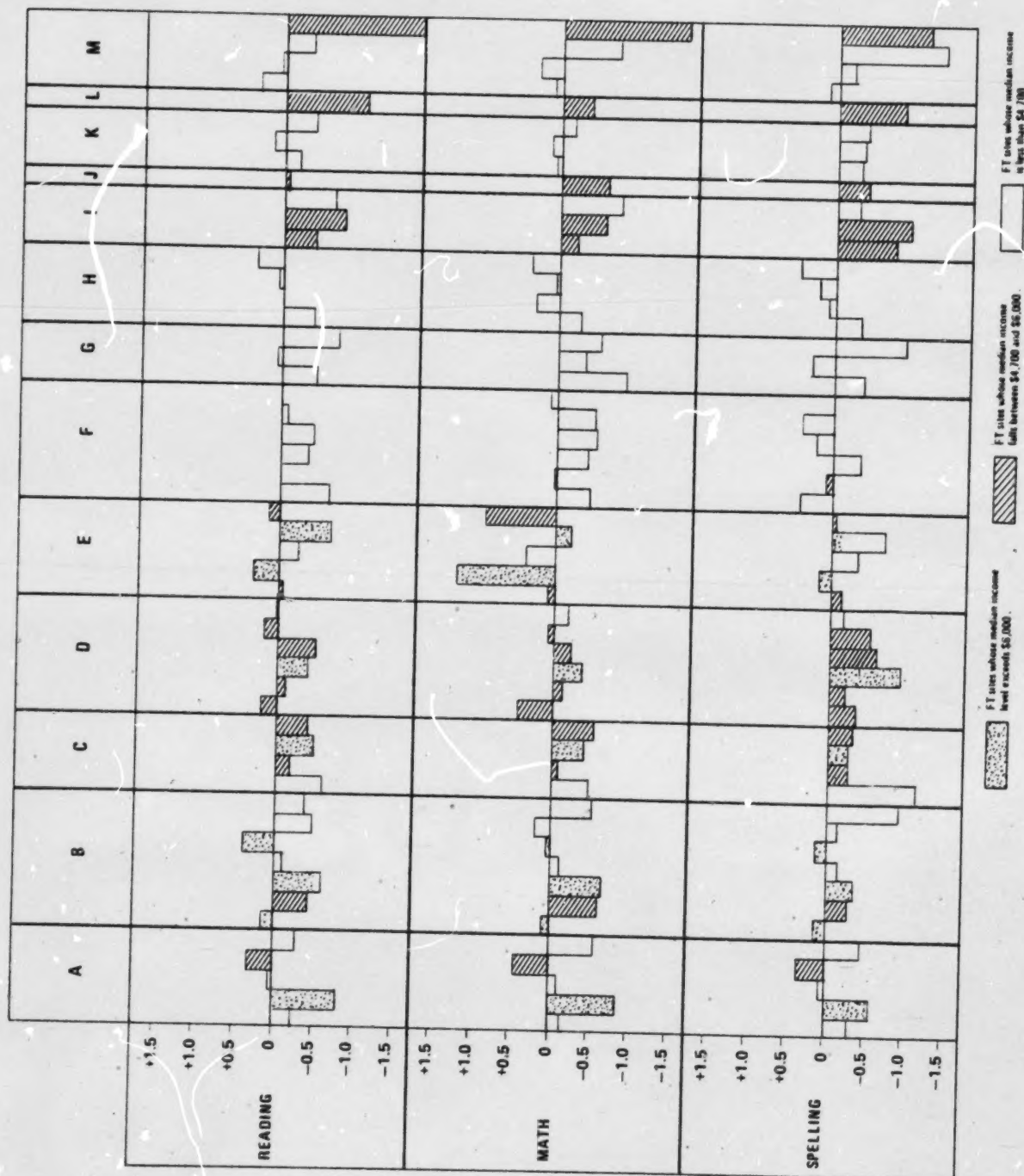


Figure 2d

DIFFERENCES BETWEEN FT AND MATCHED NFT ADJUSTED THIRD GRADE OUTCOMES
ON MAT READING, MATHEMATICS AND SPELLING IN STANDARD DEVIATION UNITS
FOR COHORT II SITES

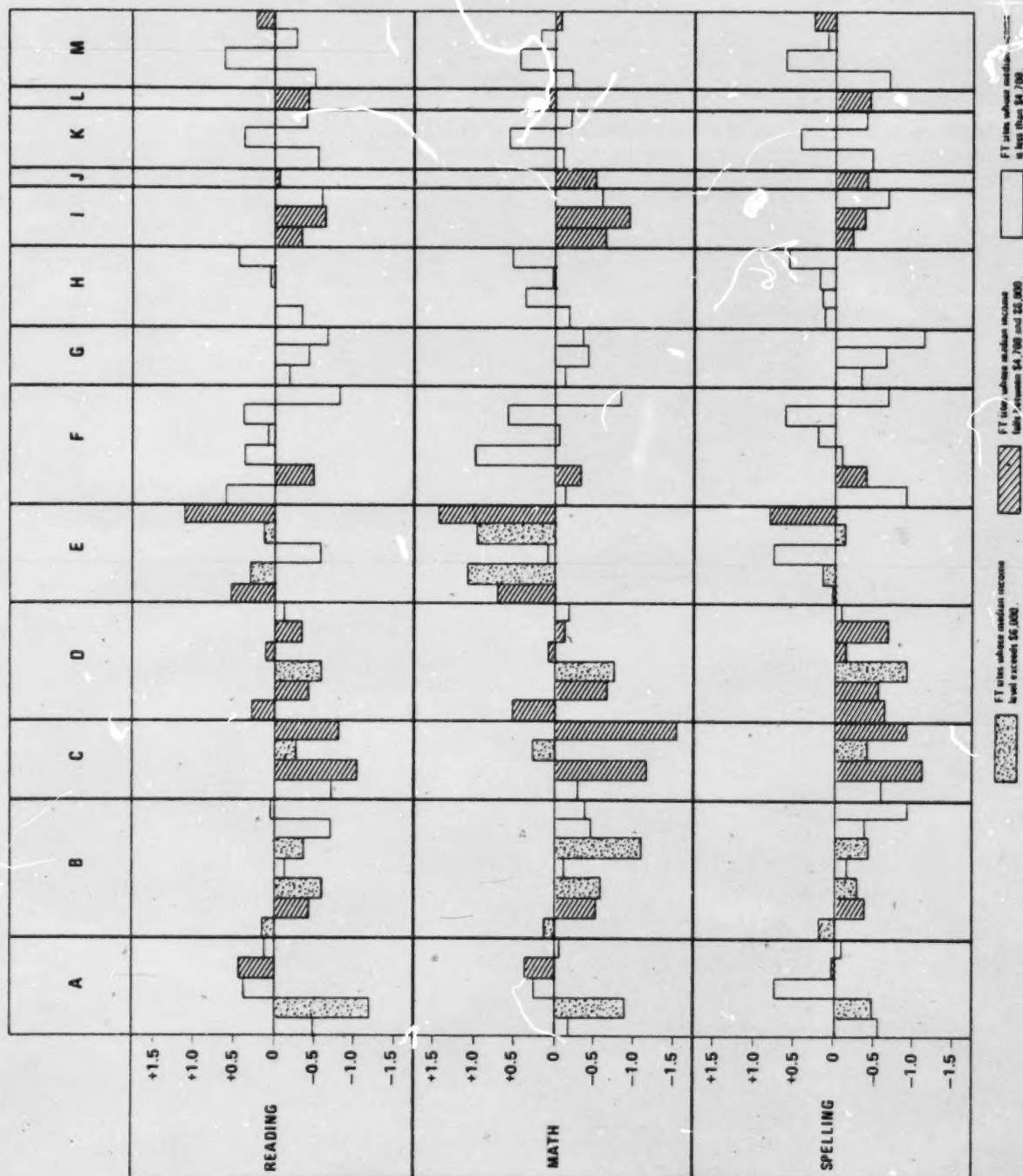


Figure 2e
DIFFERENCES BETWEEN FT AND POOLED NFT ADJUSTED THIRD GRADE OUTCOMES
ON MAT READING, MATHEMATICS AND SPELLING IN STANDARD DEVIATION UNITS
FOR COHORT II SITES

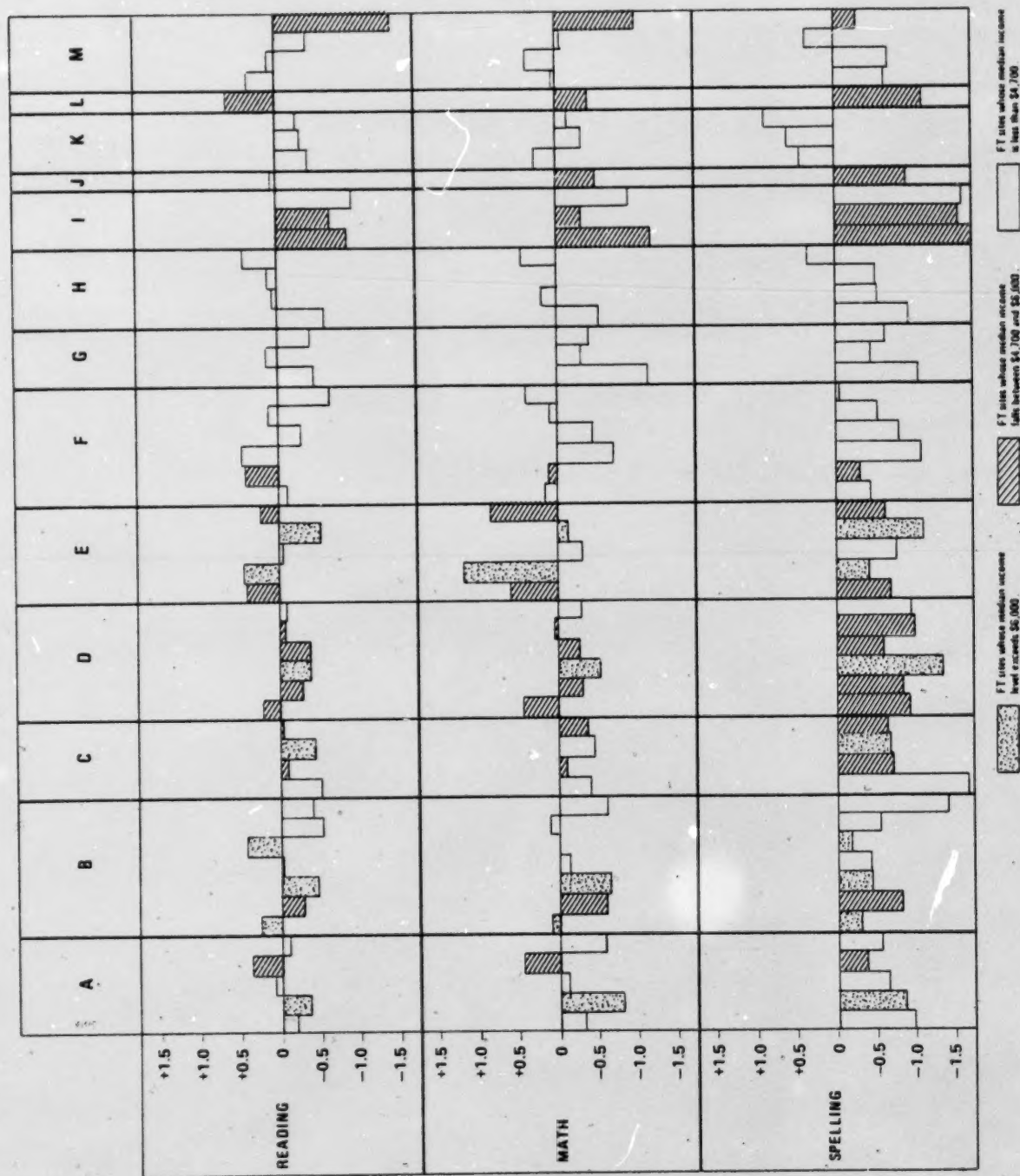
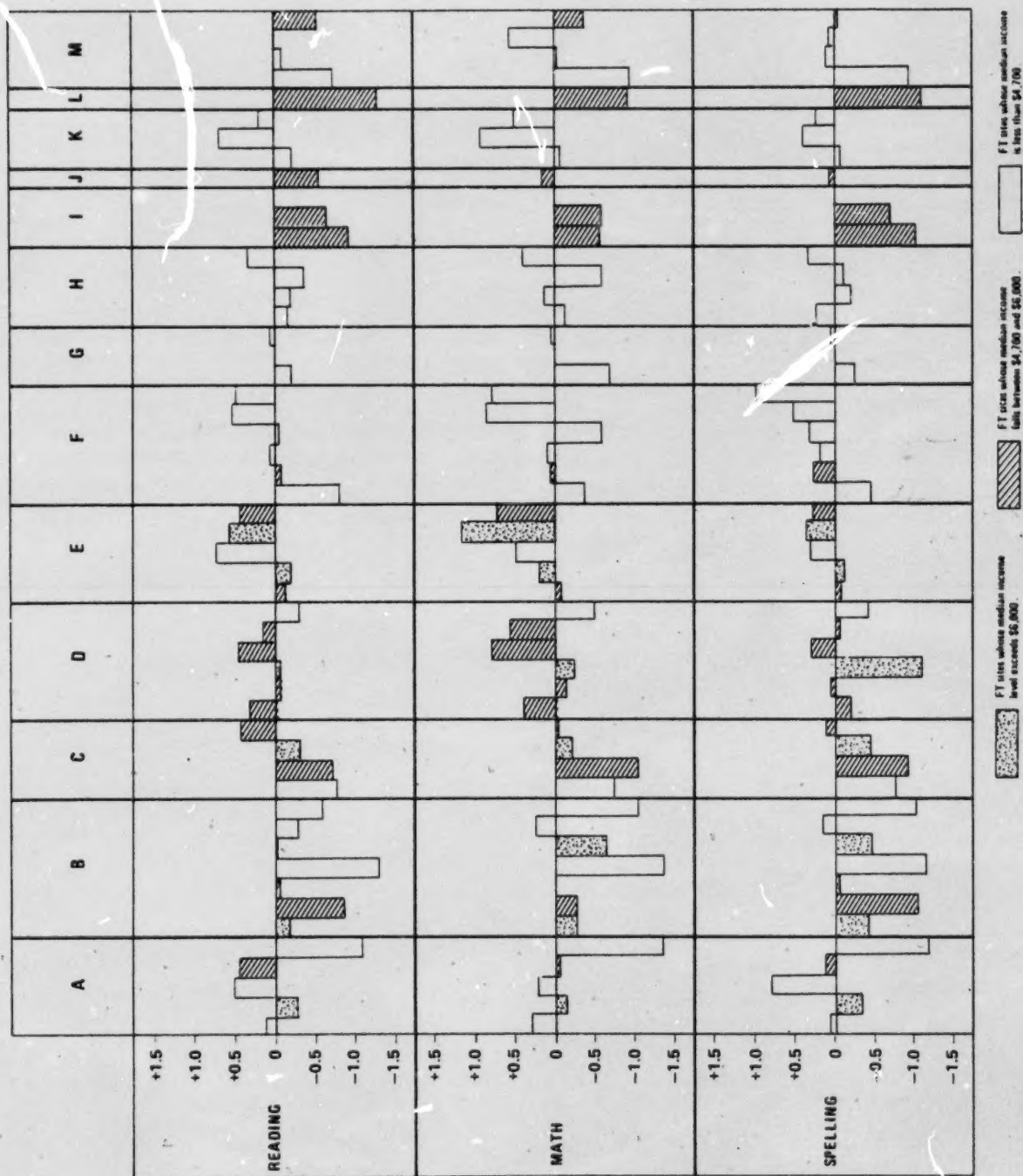


Figure 2f

CHANGES IN RANK ORDER OF FT AND NFT MEAN SCORES
FROM WRAT PRETEST TO MAT READING, MATHEMATICS AND SPELLING POSTTEST
FOR COHORT II SITES



an adequate adjustment of the outcome, and we judge that it is not quite reasonable to compare the spelling progress of English-speaking children directly with that of non-English speaking children. The pooled comparison, on the other hand, should handle the adjustment better, since it can take advantage of the correlation of the outcome with the ethnic/linguistic categories across the entire NFT population, not merely within the site in question. The pooled analysis, moreover, includes appropriate interaction terms for taking account of language differences.

It is clear that each analysis is telling the truth from its own perspective:

- The local and matched analyses say, "These Follow Through children did not learn as much spelling as did these non-Follow Through children, more of whom are native speakers of standard English."
- The normative analyses say, "Even after four years of Follow Through treatment, these Follow Through children, many of whom speak a native language other than standard English, do not spell as well as the general population."
- The pooled analysis says, "These children scored better on the spelling posttest than our regression model would lead us to expect similarly-disadvantaged NFT children to do."

In as complex a situation as the Follow Through quasi-experiment, we submit that it is valuable to have such a variety of analytic perspectives. One would not wish to hang an evaluation of Sponsor K's performance in teaching spelling exclusively on any one of the three statements above. Taken together, the three provide a plausible and balanced picture. We could always wish for more data of different kinds, especially now that we have the benefit of hindsight; but given the data we have, analytic cross-validation provides more confidence in our conclusions than any of its component analyses could by itself.

One final observation: it is difficult, expensive, tedious, and time-consuming to interpret the data from a quasi-experiment. The larger the quasi-experiment, and the further it deviates from a true experimental model, the harder is the job and the less trustworthy are the results. Follow Through, by its sheer size and complexity, illustrates these generalizations as well as any enterprise in history. Nine years ago, it would

probably not have been politically feasible to undertake the kind of randomized study in compensatory education that ESAA has since funded and in which many federal agencies are showing increased interest. Now it is more evident what one has to pay for the ease of implementation that a quasi-experimental design can provide. I trust that the experience of Follow Through has provided an important part of the motivation for this evolution in policy.

REFERENCES

- Campbell, D. T., et al. In H. W. Riecken & R. F. Boruch, Social experimentation: A method for planning and evaluating social intervention. New York: Academic Press, 1974. Pp. 38-9, 235-9, 261-2.
- Cook, T. D. The potential and limitations of secondary evaluation. In M. W. Apple, M. J. Subkoviak, & H. S. Lufler, Jr., Educational evaluation: Analysis and responsibility. Berkeley, Ca.: McCutchan Publishing Co., 1974. Pp. 168-9, 174-81, 206-7.
- Smith, M. S. Some short term effects of project Head Start: A preliminary report on the second year of planned variation. Cambridge, Mass.: Huron Institute, 1973.
- Weisberg, H. Short term cognitive effects of the Head Start program: A report on the third year of planned variation - 1971-72. Cambridge, Mass.: Huron Institute, 1974.